Beyond Concordance Lines: Using Concordances to Investigating Language Development

ARSHAD ABD. SAMAD
Faculty of Educational Studies
Universiti Putra Malaysia
E-mail: arshad@educ.upm.edu.my

Abstract

The language corpus offers a useful resource in language teaching and has been commonly been used as the basis of dictionaries and teaching materials (Woolard, 2000). Its use in language analysis is today enhanced by the availability of concordancing. In the Malaysian context, the use and analysis of language corpora has been somewhat limited. Earlier effort includes work done and compiled by researchers at Universiti Teknologi Malaysia. More recent efforts are the Malaysian learner's corpus collected in 2002 by Universiti Malaya and the English of Malaysian school students or EMAS corpus by researchers from Universiti Putra Malaysi.and consists of close to half a million words (Arshad et al., 2002). It is an untagged and unedited learner corpus that contains written data in the form of three essays written by about 800 students. The students involved were 11 year olds in Year 5 of primary school education, as well as 13 year olds in Form 1 and 16 year olds in Form 4 in the secondary school. This article examines language development based on data in the EMAS corpus using language production as well as lexical variety as indicators of development.

Introduction

In the present day and age of computer assisted language learning or CALL, the language corpus offers a useful resource in language pedagogy. The language corpus has commonly been used as the basis of dictionaries and teaching materials (Woolard, 2000). Its use is today enhanced by the easy availability of concordancing software such as Wordsmith, MonoConc Pro and Microconcord. These software help in the tedious task of analysing language data and greatly extend the potential of a corpus in language pedagogy. Corpora often help to inform on how words and grammatical constructions are used. Teachers, researchers and even language learners typically examine concordance lines to discover how words and grammatical constructions are used. Schmitt (2002:34), argues that among the benefits of using a corpus in language teaching and learning is that it may help students to "look at the systematicity of language as an interesting linguistic puzzle, rather than a set of boring rules to be memorized". New software such as RANGE, developed by practitioners in linguistics have provided an additional perspective to corpus studies (Nation, 2002). In the Malaysian context, the use and analysis of language corpora has been somewhat limited. Although there have been mention of an earlier corpus compiled by researchers at Universiti Teknologi Malaysia, to the best of this writer's knowledge, this corpus is by-and-large unavailable. Two more recent efforts are the Malaysian learner's corpus collected by Universiti Malaya and the English of Malaysian school students or EMAS corpus by researchers from Universiti Putra Malaysia. This article investigates language development based on data in the EMAS corpus using language production as well as lexical variety as indicators of development.

The EMAS Corpora

The EMAS corpus used in this study was collected in 2002 and consists of close to half a million words (Arshad et al., 2002). It is an untagged and unedited learner corpus that contains written data in the form of three essays written by about 800 students. The students involved were 11 year olds in Year 5 of primary school education, as well as 13 year olds in Form 1 and 16 year olds in Form 4 in the secondary school. All students who contributed to the corpus were considered as being above average in English language proficiency.

The first essay was based on a picture series and was administered by the researchers during visits to school. Students were given an hour to write an essay describing events that occur during a fishing trip. The second essay was an essay entitled “The happiest day of my life” which teachers of the selected schools administered to the respondents. The third essay was selected by teachers from essays that respondents had completed as part of their regular school work.

The major criterion in selecting the topics for the essays was the amount of language the topics could elicit. As such, topics and tasks such as describing a picture sequence involving an incident at a river were selected because students were expected to be able to write more because of their familiarity with the task and topic. The picture essay was also taken from a Year 6 textbook so that the task would not be too difficult for the younger respondents in the study. Additionally, other corpus initiatives such as the National Science Foundation project also use a similar pictorial or visual prompt. In order to elicit a large amount of language data, the essay topic “The happiest day of my life” was used as the topic for one of the essays as almost every respondent was thought to be able to write on the topic and hence produce the required amount of language data.

Investigating Development

Various methods can be used to determine language development. Numerous language acquisition studies, for example, focus on specific target structures and examine the acquisition of these structures over a period of time. It should be noted that the available data in the EMAS corpus is cross sectional. However, data in the corpus were elicited from three age groups. A basic assumption made in this article, therefore, is that developmental patterns can be implied by comparing the language use of the three different age groups. In this article, developmental patterns are examined by studying the language productivity as well as vocabulary use of students. Language productivity involves the amount of language produced while vocabulary use refers to the sophistication of the vocabulary as well as the use of academic vocabulary based on a language corpus. The study examines language development by comparing the performance of the three age levels with regards to their language productivity and vocabulary use.

Language Productivity

Productivity in this article is indicated by the number of sentences per essay and the words per sentence. The productivity of the three age groups is compared in order to examine language development. Table 1 presents the productivity of the students according to the indicators mentioned using the entire corpus.

Table 1: Number of Sentences Per Essay and Words Per Sentence According to Age Level

	Sentences	**Sentences/Essay**	**Words**	**Words/Sentence**
Primary 5	9,028	13.19	82,218	9.11
Form 1	10,077	14.68	118,353	11.74
Form 4	16598	31.08	201,372	12.13

We can detect a gradual increase in the number of sentences, sentences per essay and words per sentence from the Primary 5 to Form 4 levels based on the information provided in Table 1. Such a finding is not at all surprising given the increased cognitive maturity of older students. Nevertheless, it is encouraging to note that the older Form 4 students actually produce longer essays, as well as longer and more complex sentences. A chi-square analysis reveals a significant increase in the number of sentences per essay according to the age of the respondents ($\square^2$ = 10.03, $p < .05$).

Range of Vocabulary

The diversity of the vocabulary used in a corpus is often determined by calculating the type to token ratio (Schmitt, 2002). This ratio is calculated by dividing the number of separate words in a text (type) by the number of words in the text (token) as in the following formula by Laufer and Nation (1995).

$$\frac{\text{Number of separate words (types)}}{\text{Number of words in a text (tokens)}} \quad \text{x } 100$$

A larger type to token ratio is interpreted as an indication of a wider range of language used. Uncommon words tend not to be used frequently. Therefore, with possible values of 0 to 100, a higher type to token ratio suggests that the learners are using many uncommon words. Lower ratios, on the other hand, may indicate an over-reliance on a limited set of words. The type to token ratio of the three age groups based on the sub corpus of picture essays in the EMAS corpus is presented in Table 2.

Table 2: Type to Token Ratio of the Three Age Groups

	Types	**Token**	**Type/Token Ratio**
Primary 5 1213 32454* 3.74			
Form 1	1773	43737	4.05
Form 4	3552	78509*	4.52

** Estimates based on average frequency of occurrence of each unaccepted word in the Form 1 sub corpus*

Table 2 indicates that the type to token ratio gradually increases from the lower to higher age groups. This signifies that the older respondents use a wider range of vocabulary in their essays. Despite the increase in ratio from the lower to the higher age groups, the observed ratio of 3.74 and above is rather low. Schmitt (2002) cites work that consider a ratio of 36 to 57 as normal for written texts.

A possible reason for the low type to token ratio is the nature of the written text itself. As narration based on a picture sequence may require constant reference to objects in the sequence, this may partially contribute to the low ratios obtained. A second more directly relevant reason is the nature of the data itself. Because the sample consists of respondents who are writing on the same essay topic, the type to token ration will be inadvertently restricted. Very often, the same words are used by many respondents in writing the essay which is based on the same picture series. As such, a new ratio is obtained that estimates the number of type as well as token per student. Assuming that the tokens produced by all respondents are more or less equal, the number of tokens in Table 2 is divided by the number of respondents for each level. The average amount of type was calculated by first counting the number of words with a frequency higher than or equal to the number of respondents in each age level. There were 14, 27 and 55 such words for the Primary 5, Form 1 and Form 4 students respectively. These amounts were considered as part of the amount of word types produced by each student for the three age groups. These word types were high frequency words such as *the*, *he*, *river* and *girls* which were all relevant to the picture sequence. In order to estimate the number of types per respondent, these amounts were also deducted from the total number of types found in Table 2 and the result was divided by the number of respondents for each age group and multiplied by three. The result of this operation was then added to the number of high frequency words found earlier to provide an estimate of the average number of word types per student. Using these modifications, the modified type to token ratios for the three age groups are presented as follows in Table 3:

Table 3: Estimated Type to Token Ratio Per Student the Three Age Groups

	Number of Respondents (N)	**Types (T)**	**Average Types**	**Tokens (Tk)**	**Average Tokens (Tk/N)**	**Average Type/Token Ratio**
Primary 5	294	1213	26.3	32454	110	23.91
Form 1	301	1773	44.4	43737	145	30.62
Form 4	264	3552	94.6	78509	297	31.86

While the type to token ratios found in Table 3 seem more acceptable, they remain conservative estimates with deflated values. Nevertheless, for the purpose of this article, it is sufficient to note that the values continue to show an increase in higher age groups.

It should be also noted that the EMAS corpus is a learner corpus that retained the students' spelling and grammatical errors. In calculating the type to token ratio, nonsensical words were deleted from the data. Similarly, proper names were also excluded. Misspelled words were corrected and counted as part of the data. These steps were also taken in examining the sophistication of the vocabulary used as described in the following section.

Sophistication of Vocabulary

The sophistication of the vocabulary can be determined by using specialized software such as RANGE (Nation, 2002), a vocabulary analysis program which gives an indication of the kind of vocabulary used. The program analyses text by comparing it to several base lists of frequently used words. The first base list includes the most frequent 1,000 words in English. The second includes the second most frequent 1,000 words, and the third includes words not in the first 2,000 words in the two previous lists but are frequent in upper secondary and university levels from a wide range of subjects. All three base lists include the base form of words and derived forms. As such, the first list of 1,000 words consists of around 4,000 forms or types. According to Nation (2002), the sources of the lists are A General Service List of English Words by Michael West (1953) for the first 2,000 words and the American Word List by Coxhead (2000) containing 570 word families.

Based on the EMAS sub-corpus of picture essays, the frequency and percentage of word types and families are presented in Table 4.

Table 4: Frequency and Percentage of Word Type and Families According to Age Groups and Categories

	Type		**Families**
	f	**%**	
Category one (Most frequent 1,000 English words)			
Primary 5	770	63.5	470
Form 1	1027	57.9	590
Form 4	1580	44.5	758
Category two (Most frequent 1,000 – 2,000 English words)			
Primary 5	250	20.6	180
Form 1	410	23.1	281
Form 4	789	22.2	492
Category three (Frequent academic words not in categories one and two)			
Primary 5	29	2.4	25
Form 1	43	2.4	40
Form 4	184	5.2	134
Unlisted category (Words not in categories one, two and three)			
Primary 5	164	13.5	
Form 1	293	16.5	
Form 4	999	28.1	
Total			
Primary 5	1213	100	675
Form 1	1773	100	911
Form 4	3552	100	1384

The frequencies and percentages in Table 4 indicate a clear development in the sophistication of vocabulary used. This development is apparent if we were to make two simple observations of the figures in the table.

Firstly, a large majority of words used by the Primary 5 students (63.5%) is found in the first category of most frequent 1,000 English words. While this observation is true of both the Form 1 and Form 4 respondents, the percentage majority is relatively smaller at 57.9% and 44.5% respectively. This indicates that not only do the older age groups tend to use a wider range of words, the words they use are also more sophisticated. Secondly, the percentage of words used by Form 4 respondents in the academic base list (5.2%) and in the unlisted category (28.1%) is higher than the other two age groups. While words in the unlisted category such as *river*, *flow* and *daisies* may be due to the topic of the specific writing task, the relatively higher percentage of words from the academic base list used is indicative of a more academic familiarity among the Form 4 respondents.

Words in the Academic base list or the Third Category used by various age levels are presented in Table 5.

Table 5: Academic Words Used by Two or More Age Levels

Words in category three used by all three age levels

aid, appreciate, couple, final, project, recover, seek

Words used by at least two groups

Primary 5 and Form 1:
assemble, ignorant, incidence, injure

Primary 5 and form 4:
accompany, job, minor, normal, occupy, participate, partner, site

Forms 1 and 4:
area, collapse, concentrate, confirm, contact, cycle, equip, eventual, identify, locate, medical, primary, principal, process, react, relax, stable, survive

Words in the academic base list used exclusively by either primary five, Form 1 or Form 4 students are presented in Appendix 1. Clearly, a more diverse list of words is used by the older Form 4 respondents.

Conclusion

A rather obvious concern in the use of a corpus in language teaching and learning is how it can actually be used. Although concordance software can be used to help analyze the language data available, many may still be easily disturbed by the huge amount of data available. This article has attempted to present the relevance of corpus data in investigating language development without having to analyze concordance lines. Additionally, it has used a local learner corpus in order to highlight the existence of such corpora as well as the efficacy of these corpora in corpus based studies.

Based on written data of three age groups, the results of the concordance and analysis indicate some form of development in terms of language production as well as vocabulary range and sophistication. Equally important, the article demonstrates that corpus data can be analyzed to inform language educators of language development. While development in this study is examined cross-sectionally, it may be possible to introduce a longitudinal element in future analyses. The data used in this study was taken from the EMAS corpus collected in 2002. Similar data can be collected from respondents of the same age group in later years and compared against the values found in this study. As Hunston (2002:206) notes, "the essence of work on learner corpora is comparison". The values in this study, therefore, can be regarded as benchmarks against which to compare future groups of students as well as assess the development of the language program in Malaysia in general.

References

Arshad Abd.Samad, Fauziah Hassan, Jayakaran Mukundan, Ghazali Kamarudin, Sharifah Zainab Syd Abd. Rahman, Juridah Md. Rashid & Malachi Edwin Vethamani (2002). *The English of Malaysian School Students (EMAS) Corpus*. Serdang: Universiti Putra Malaysia.

Coxhead, A. (2000). *A New Academic Word List. TESOL Quarterly*, 34: 213-238.

Laufer, B. & Nation, I. S. P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics,* 16(4): 307-322.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: CUP.

Nation, P. (2002). *Range and Frequency: Programs for Windows Based PCs*.

Schmitt, N. (2002). Using Corpora to Teach and Assess Vocabulary. In Tan, M. (Ed.). *Corpus Studies in Language Education* (pp. 31-44). Thailand: IELE Press.

Woolard, G. (2000). Collocation: Encouraging Learner Independence. In Lewis, M. (Ed.). *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 28-46). Hove, England: Language Teaching Publications.

Appendix 1: Words in the Academic Category used exclusively by each age group

Form 4

abandoned, academic, accompanied, achieve, alter, annual, anticipating, apparently, appreciated, appreciation, approached, approaching, appropriate, approximately, aspects, assembled, assembly ,assign, assignments, assist, assume, assumed, assure ,assured, automatically, available, aware, behalf, challenged, challenging, coincidence, collapsed, collapsing, commented, compound, consists, constant, constantly, consult, convince, convinced, cooperated, cooperating, cooperation, couples, create , creating, creation, creativity, credits, cycled, cycles, cycling, declined, deduction, definitely, depressed, design, despite, distinction, energy, ensure, environment, equipment, equipped, eroded, estate, estimated, eventually, expert, factor, filed, finally, focus, fund, hence, identified, ignore, ignored, ignoring, incident, incidentally, injured ,injuries, injury, instruction, instructor, intelligent, internal, involved, isolated, issues, items, journal, labeled, labouring, located, location, logical, method, methods, migrated, monitored, normally, obvious, obviously, occupied, occurred, participated, physically, potential, predict ,predicts, previous, proceeded, professional, ranging, reacted, reacting, recovered, relaxation, relaxed, relaxing, release, released, relied, reluctant, rely, remove, resident, responded, response, revision, role, schedule, seeking, series, shifted, similar, sites, somewhat, stabilize, strategic, stress, submit, sum, survey, survived, sustain, task, team, tension, topic, unaware, unique, unstable, utilities, vision, volume

Form 1

apparent, coincide, definite, intense, medium, occur, reject, respond, restore, strategy, technique

Primary 5

accommodate, adult, founded, grant, maintain, pose